Repositories and the Cloud

The Cloud—cloud computing, cloud storage, cloud software has become a hot topic over the last year, filled with much hype and promise. However, cloud computing has the potential to transform a large part of the IT industry, making software even more attractive as a service and shaping the way IT hardware is designed and purchased. The elasticity of resources, without paying a premium for large scale, is unprecedented in the history of IT.* The Gartner group defines the "Cloud" as a style of computing where massively scalable IT-related capabilities are provided “as a service” using Internet technologies to multiple external customers. This panel will discuss the efforts of three distinctive projects that are making cloud services accessible and integrated with repository platforms. The speakers and topics are as follows:

Michele Kimpton and Sandy Payette: DuraSpace

The DSpace Foundation and Fedora Commons are investigating the feasibility and interest of a new service named DuraSpace to serve academic libraries, universities, and other organizations in providing perpetual access to digital content. DuraSpace can be understood as a Web-based service that makes stored digital content more durable, manageable, accessible, and easier to share. A key design feature of DuraSpace is to leave the basics of pure storage those who do it best (storage providers) and to overlay storage solutions with additional functionality that is essential to ensuring long-term access and ease of use. The service provides baseline functionality that begins with the ability to replicate and distribute content across multiple cloud providers. It adds value over and above storage by enabling the deployment of services to support access, preservation, re-use, and sharing of content stored in the cloud.

Richard Rodgers: Cloud Task Replica - Towards a Preservation Strategy

DSpace has always described it's mission in terms of long-term preservation of digital assets, and provides some architectural support (e.g. in strong typing of asset formats) for constructing and executing preservation strategies on its content. However, comparatively little work has been done on ensuring content viability against catastrophic loss, which is a precondition for all other preservation functions.

In the digital realm, replication is typically the keystone of such a strategy, and its general principles are reasonably well understood. To be effective, a replication management system must go substantially beyond standard IT backup practices, which protect only against isolated, local, technical or procedural failures. This high-level objective raises a host of issues, among them:

* How to define and package an asset in its entirety.

* How to establish replication partnerships and relationships with other entities that can be effective, auditable, but low-cost.

* How to manage the actual mechanics of replication in a flexible, scalable, and resource-effective manner.

This presentation will share our experience in the design of a replication system whose focus is largely in this last area, but which
might also begin to shed some light on other aspects of the problem. The design attempts to model the problem in a manner flexible enough so that replication can be entirely self-managed within a repository, by a network of peer institutions, or service providers, or any combination of the above, or indeed evolve from one to another. It also envisions replication relationships among heterogeneous repository systems, e.g. DSpace and Fedora.

Dave Tarrant, Tim Brody, Les Carr: From the Desktop to the Cloud: Leveraging Hybrid Storage Architectures in Your Repository

Repositories collect and manage data holdings using a storage device. Mainly this has been a local file system, but recently attempts have been made at using open storage products and cloud storage solutions, such as Sun's Honeycomb and Amazon S3 respectively. Each of these solutions has their own pros and cons but There are advantages in adopting a hybrid model for repository storage, combining the relative strengths of each one in a policy-determined model. In this paper we present an implementation of a repository storage layer which can dynamically handle and manage a hybrid storage system.

*Above the Clouds: A Berkeley View of Cloud Computing, Michael Armbrust et al, 2/2009